AMENDMENT TO THE CLAIMS

1-16 (Canceled)

17. (Currently Amended) A method of developing a corpus for training a language model implemented by a computer, comprising:

extracting a list of potential words from a first corpus of text that match defined words and rules;

~~determining if the list includes a sufficient number of defined words and rules;~~

accessing human ~~annotating~~ annotations of the first corpus ~~to~~ that provide indications of word type for the text in the first corpus; ~~and~~

providing morphological tags in the first corpus indicating a morphological type of an associated sequence of characters and a combination of parts forming a morphological subtype as a function of the annotations; and~~.~~

annotating a second corpus of text that is larger than the first corpus with indications of word type for the text in the second corpus using the computer based on the first corpus.

18. (Currently Amended) The method of claim ~~15~~ 17 wherein ~~annotating further comprises providing~~the annotations comprise indications of whether the word is in a lexicon, a morphologically derived word, a factoid and a named entity.

19. (Original) The method of claim 17 wherein the morphological type is one of affixation, reduplication split, merge and head particle.

20. (Original) The method of claim 17 wherein providing morphological tags further comprises indicating a part of speech for the combination of parts.

21. (Original) The method of claim 17 wherein providing morphological tags further comprises indicating a pattern of characters for the combination of parts.

22. (Canceled)

23. (New) The method of claim 17 and further comprising:

using the second corpus to train a language model; and

processing a test corpus of text using the language model to provide an output indicative of words in the text of the test corpus and an indication of word segmentation of the text in the test corpus.

24. (New) The method of claim 23 and further comprising:

comparing the output of the language model with a predefined annotation defining word segmentation for the test corpus; and

evaluating the output of the language model based on the comparison.

25. (New) The method of claim 24 wherein comparing includes identifying matching words that appear in both the output and the predefined annotation.

26. (New) The method of claim 25 wherein evaluating provides a value of effectiveness based on words that match.

27. (New) A method for evaluating a word segmentation language model, comprising:

building the word segmentation language model based on an annotated corpus;

applying the language model to a test corpus of unsegmented text different from the annotated corpus to provide an output indicative of words in the test corpus;

comparing the output of the language model with a predefined word segmentation of words of the test corpus; and

evaluating the language model based on the comparison of the output and the predefined word segmentation to provide an indication of effectiveness of the language model as a function of individual types of words.

28. (New) The method of claim 27 wherein evaluating further comprises identifying words in the output that match words in the predefined word segmentation.

29. (New) The method of claim 27 wherein comparing comprises comparing person names, location names, organization names, overlapping ambiguous strings and covering ambiguous strings in the output and the predefined word segmentation.

30. (New) The method of claim 29 wherein the indication of effectiveness is contacted based on only the comparison of person names, location names, organization names, overlapping ambiguous strings and covering ambiguous strings.